Using a Proximity Filter to Improve Rabies Surveillance Data

Andrew Curtis*
Louisiana State University, Baton Rouge, LA

Abstract

The proximity filter is one of the new exploratory spatial data analysis techniques developed as a result of the visual and interactive capabilities of geographic information systems (GIS). This technique, a variant of the spatial filter, is used to identify significant "holes" in a point surface. This paper presents an example of how the proximity filter can be applied to real-world situations, showing how it can improve rabies surveillance data. The points used in this example are the locations of animals submitted for rabies testing. It is important to identify any holes in a rabies surveillance data surface to see if the low occurrence of points (i.e., of animals submitted) results from natural reasons or from a lack of local education. The proximity filter compares the number of points in an inner ring to those in an outer ring. A Monte Carlo simulation allows for a test of significance to be constructed. A GIS is used to vary the size and shape of the inner and outer rings used in the analysis. Two types of filter shape are discussed; the first is circular, and the second uses the buffer function of the GIS to mirror the shape of an investigated area (such as a county). This paper also provides an example of how data aggregated to a political unit can be turned into a simulated point pattern surface so that the proximity filter can be applied.

Keywords: exploratory spatial data analysis, modifiable areal unit problem, Monte Carlo, point pattern analysis, spatial filter

Background

Recent discussion has focused on the role geographic information systems (GIS) can play in improving spatial analysis (1). One area in which such advancement can be made is exploratory spatial data analysis—the use of a GIS' interactive and visual capabilities to search for problem solutions on the fly (2). Although this practice is still in its infancy, new techniques of analysis that can only be implemented within a GIS environment are now being developed (for a review of these techniques, see reference 3). A good example of these techniques is the spatial filter, a form of analysis that avoids the problems associated with data aggregated to political units (county, zip code, census tract, etc.). Diseases are often continuous across space, and therefore, if the data are available at a disaggregated level, it makes little sense to impose artificial areas (defined by political boundaries) on the analysis. An additional problem in using aggregated data is one of aggregation (scale) and shape (zone)—otherwise known as the modifiable areal unit problem. It has been found that different results can occur depending on which scale and zone scheme is chosen (4).

* Andrew Curtis, Louisiana State University, 110 Howe/Russell Geoscience Complex, Baton Rouge, LA 70803 USA; (p) 225-388-6198; (f) 225-388-4420; E-mail: acurti1@lsu.edu

The spatial filter avoids the modifiable areal unit problem by using original point data (5,6). A fine grid is layered onto the research surface, with the intersection points of the grid acting as foci for a series of overlapping “filters” that are used to calculate “rates” from the point data. These rates, when assigned to the intersection point, can then be mapped, usually as a contour surface. This creates a visual impression of clusters from a continuous surface. Different-sized filters will “smooth” the data to different degrees, but the most dramatic cluster groupings should emerge irrespective of filter size. This form of analysis is truly “exploratory” because the visualization of the data drives the analysis. A test of significance can then be conducted using a Monte Carlo simulation. In such a test, the original points are assigned probabilities of occurrence (such as the chance that a particular birth will become a death). The same filter analysis is then performed on the simulation surface. Through multiple repetitions of this simulation, a distribution can be created against which the original clusters can be compared to see their probability of occurrence.

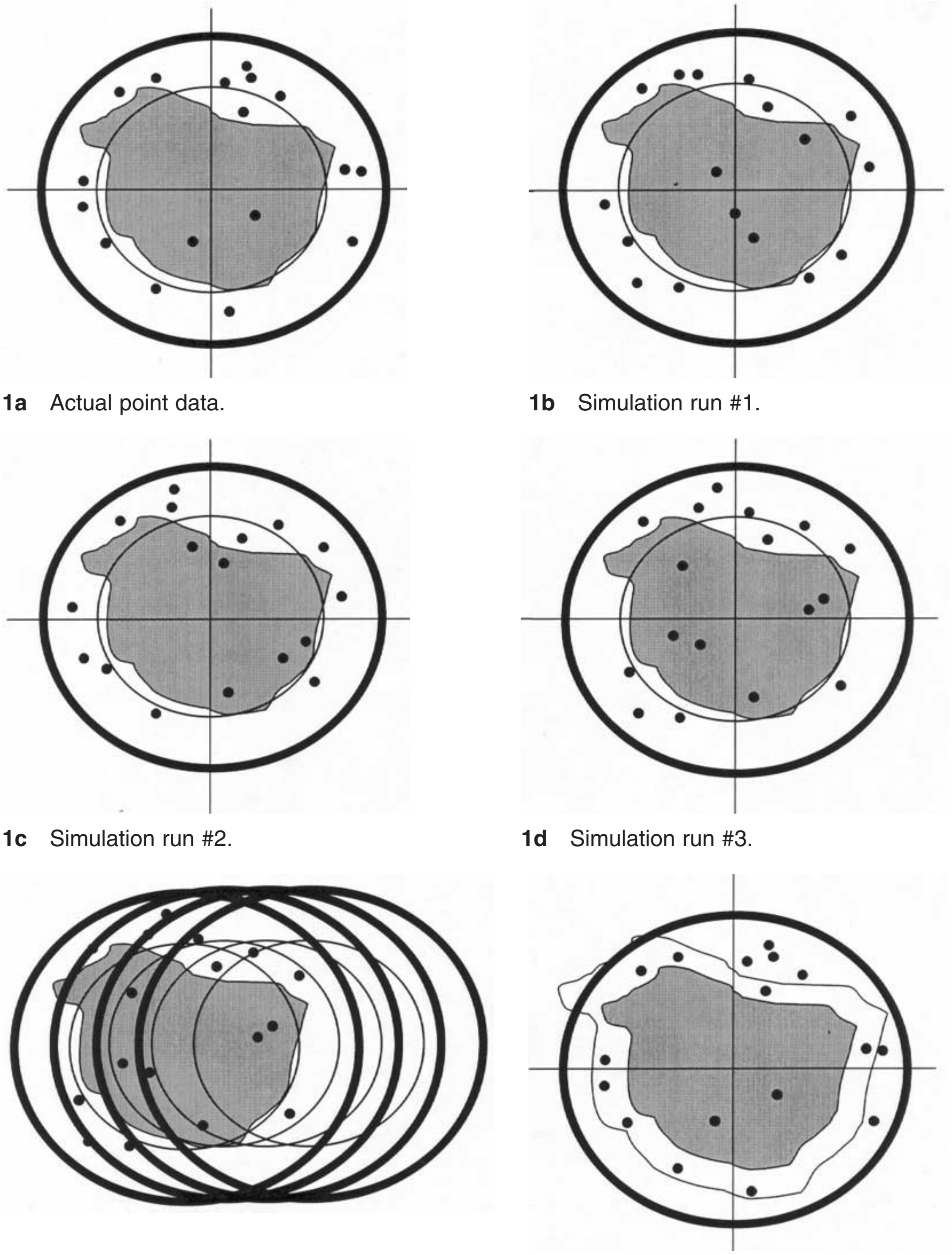
Although the spatial filter is useful for identifying significant clusters on a point surface, the research problem for this paper required a technique that could find significant “holes” in a point surface. A variation of the spatial filter was needed—the proximity filter, which can use a GIS to compare two distributions of points and identify significant differences between them.

Rabies Surveillance Data

A raccoon rabies epizootic has been spreading through the eastern seaboard states since 1977. The epizootic started when rabid raccoons were translocated to West Virginia/Virginia by hunters (7). Since then, most of the eastern seaboard states have been affected by the spread (8). In a state impacted by rabies, surveillance data provide the main source of information about the disease. Both the public and local officials are supposed to submit animals for rabies testing if they interact with people (i.e., bite or scratch them) and cannot be quarantined, or if they display potential rabies symptoms. Data from this testing are used to determine how far the disease has progressed in the state and where to place countermeasures such as oral vaccine barriers. These data are also used to determine the success of these countermeasures. Unfortunately, evidence suggests that data quality varies considerably between areas (9). There is, therefore, a need for a means of analysis that can identify areas where fewer animals are submitted for rabies testing so that resources can be targeted to “educate” both the public and local officials.

The Proximity Filter

The general principle of the proximity filter is the same as that of the spatial filter. It uses a floating kernel to analyze a point data surface. The main difference between the proximity and spatial filters is that the former compares an inner and outer spatial point distribution. Figure 1a shows a series of points (the locations of animals submitted for rabies testing), the boundary of the county under investigation, and an inner and outer ring of the spatial filter (centered on the cross hair). In this example, there are 17 points in the inner and outer rings combined. The number of points in the inner ring (in this case, three) can be compared with the number in the outer ring (in this case, 14) to



1a Actual point data.

1b Simulation run #1.

1c Simulation run #2.

1d Simulation run #3.

1e Overlapping proximity filter analysis.

1f Using the GIS buffer function to create the outer ring.

Figure 1 Applying the proximity filter to point data.

determine if, given the latter number, there are significantly fewer points in the inner ring than expected. GIS allows us to answer this question with a Monte Carlo simulation. A fine grid of coordinates is layered over both the inner and outer rings. All the points are then randomly redistributed across the combined area, with each coordinate having an equal chance of accepting one point. This simulation is repeated 100 times. If three or fewer points occur in the inner ring on no more than five of the simulation runs, then there is (at least) a 95% chance that the low number of points in the inner ring did not occur by chance. Figures 1b, 1c, and 1d show three typical simulation runs, with five, six, and seven points being placed in the inner circle.

The size of the outer ring obviously affects the number of points being compared with the inner ring—the larger the outer ring, the more points will be included in the analysis. One possible general rule is to use a comparable land area in the analysis; that is, the land area of the outer and inner rings should be the same. In order to avoid any bias in selecting the centroids of the filters, overlapping proximity filters should cover the entire area (Figure 1e) as in traditional spatial filter analysis. The rings should be centered on a fine grid of coordinates. Repeating the same simulation procedure for each inner and outer ring makes it possible to identify significant data holes across the entire surface.

A further modification of the proximity filter could involve using the shape of the county as the inner ring. If it is believed that this particular political boundary does exert some influence on the analysis—one county, for example, may have officials who are less compliant with submission laws—then it may be useful to include the shape of the county in the analysis. One way to do this is to leave the outer ring the same, but use the shape of the county as the inner ring. This produces changes like those shown in Figure 1f (using the same original point distribution as in Figure 1a), in which there are 2 points in the inner ring and 15 in the outer ring. (Note: Larger county-shaped outline not applicable here.) The simulation procedure described above would be used to create a test of significance.

If the shapes of the counties vary considerably (e.g., if some are compact and others elongated), then the buffer function of a GIS can be used to mirror the shape of the investigated county, making the outer ring a projected extension of the county shape. Figure 1f (mentioned above as showing a county-shaped inner ring with a circular outer ring) is also an example of this method. In it, there are nine points in the county-shaped outer ring. Again, the size of the buffer affects the number of points in the outer distribution; a possible general rule is to use the same land area for both buffer and investigated county.

The size and shape of the proximity filter depends on the background of the analysis. With no prior reason to suspect a particular county of low compliance, a complete coverage of overlapping filters is best. In the case of animals submitted for rabies testing, if there is reason to suspect a political unit such as the county, then the shape of the political unit could be used as the inner ring of the proximity filter, with either a circular or a county-shaped buffer as the outer ring.

It must be stressed that the proximity filter only works if combined with local expertise. Once a significant point hole has been identified, the next stage of the analysis is to investigate causative factors. These can include a difference in terrain between the inner and outer rings (the most extreme example being that no animal submissions are coming from the inner ring because it is a lake). A difference in human populations (the fewer the people, the less likely the human/animal interaction) or a different environment leading to human/animal interaction (e.g., the outer proximity filter containing a state park) could also explain away significantly low point counts for the inner ring. All of these factors can be investigated after the initial analysis, and some of the holes discounted accordingly.

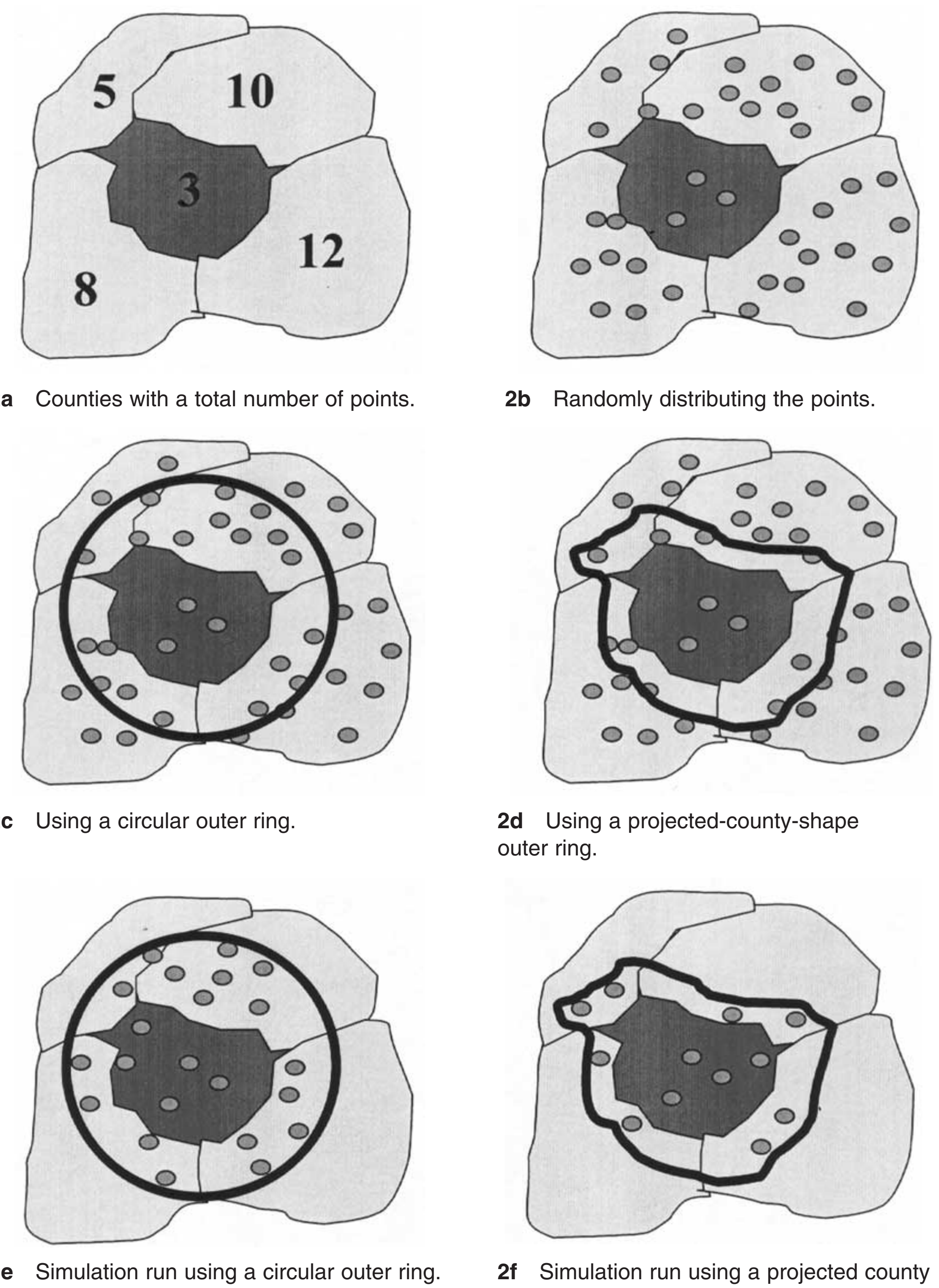
Using a Proximity Filter on Aggregated Data

Unfortunately, not all states have precise spatial locations allocated to animals submitted for testing. This is the case for Kentucky, where these data are aggregated to the county from which the submission came. Obviously, this makes it difficult to apply the filter. The only solution for cases such as this is to distribute the number of animal submissions across the host county at random.

This is easy to do using a GIS. A fine grid of coordinates is layered over the county. All the points (animals submitted for testing) are then randomly distributed across this surface, with each coordinate having an equal chance of accepting a point. Once the entire surface has been covered in this way, the proximity filter can be layered on top of "suspect" counties. There is little point in applying an overlapping filter analysis because the lack of precise locations already produces error.

Figure 2 shows the typical steps of this type of analysis. The aggregated total number of animals submitted for rabies testing is randomly distributed across the county space as a set of points, using a fine lattice of coordinates in the GIS (Figure 2b). The county under investigation (which has three points in this case) is then investigated to see if the number of submissions is significantly low given submission numbers from the surrounding counties. Either a circular or a county-shaped proximity filter can be used in the analysis of the county under investigation. In this case, the circular proximity filter contains 22 points (Figure 2c), and the county-shaped buffered proximity filter contains 12 points (Figure 2d). When dealing with aggregated data and a simulated surface, it is preferable to include the shape of the county as both the inner and outer rings of the proximity filter. This ensures that the same distance extends from the boundary of the county in all directions, reducing variations in the analysis that the county shape might cause.

The proximity filter is layered on the randomly distributed points and the points falling inside the inner and outer rings are counted. (The initial simulation generating the points from the aggregated data should be repeated several times in order to obtain an average number of points in the buffer area.) If the inner ring of the proximity filter is the county shape, then the number of points in it will always equal the total points submitted from that county. All the points falling within the inner and outer ring are then randomly redistributed across the same area, using the same fine, layered grid of coordinates within the GIS. Two examples of this simulated run can be found in Figure 2e (for the circular outer ring) and Figure 2f (for the buffer-shaped outer ring). This simulation is repeated 100 times. The distribution of points falling within the county under investigation is then compared with the simulation runs. If there is a significantly low number of points in the inner ring (investigated county) on no more than five of the simulation runs, then there is (at least) a 95% chance that the low number of points in the inner ring did not occur by chance.



2a Counties with a total number of points.

2b Randomly distributing the points.

2c Using a circular outer ring.

2d Using a projected-county-shape outer ring.

2e Simulation run using a circular outer ring.

2f Simulation run using a projected county shape.

Figure 2 Creating a point surface from aggregated data.

Improvements

Although it has been stated that the presence of data holes should stimulate a search for factors leading to their explanation, it is possible that some of these factors could be considered in the initial analysis. During the simulation, certain terrain may be more likely to accept a point than other terrain (e.g., suburban fringe compared to marshland). Using GIS makes it possible to identify land cover and create a probability surface of point acceptance. However, the question remains whether adding this complexity to the initial analysis is more efficient than investigating the holes in a homogeneous surface.

It is also useful to repeat the analysis temporally to see if there are any changes in the pattern. If the analysis identifies a county as a significant hole every year, then there is a definite reason for that hole (a difference in terrain, a problem with local officials, etc.). If a county is identified as a significant hole for only certain years, then terrain is unlikely to be the cause. An even more disturbing result would be if there were a clear temporal break point, with a series of normal years being replaced by a series of significant holes. This would suggest that something had changed in the reporting environment of the county.

The proximity filter is a relatively easy technique that can be performed using a PC-based GIS. It enables a health official to investigate any point-data source for which the presence of a hole is as important as that of a cluster. Although the example given here is for animals submitted for rabies testing, any reportable disease with an environmental association (Lyme disease, histoplasmosis, etc.) could be investigated in the same way.

Acknowledgments

Andrew Curtis would like to thank MB Auslander, DVM, MSPH, for his help at the Kentucky Department for Health Services; Ron Mitchelson for discussions during manuscript preparation; and Morehead State University (Kentucky) for its internal grant supporting this project.

References

1. Anselin L. 1999. Interactive techniques and exploratory spatial data analysis. In: *Geographical information systems: Principles, techniques, applications and management.* Vol. 2. Ed. PA Longley, MF Goodchild, DJ Maguire, DW Rhind. New York: John Wiley & Sons. 253–66.
2. Openshaw S, Alvanides S. 1999. Applying geocomputation to the analysis of spatial distributions. In: *Geographical information systems: Principles, techniques, applications and management.* Vol. 2. Ed. PA Longley, MF Goodchild, DJ Maguire, DW Rhind. New York: John Wiley & Sons. 267–82.
3. Gatrell A, Senior M. 1999. Health and health care applications. In: *Geographical information systems: Principles, techniques, applications and management.* Vol. 2. Ed. PA Longley, MF Goodchild, DJ Maguire, DW Rhind. New York: John Wiley & Sons. 925–38.
4. Curtis A, MacPherson AD. 1996. The zone definition problem in survey research: An empirical example from New York state. *The Professional Geographer* 48:310–20.

5. Openshaw S, Charlton M, Craft AW, Birch JM. 1998. Investigations of leukemia clusters by the use of a geographical analysis machine. *The Lancet* I:272–3.

6. Rushton G, Lolonis P. 1996. Exploratory spatial analysis of birth defect rates in an urban population. *Statistics in Medicine* 15:717–26.

7. Jenkins SR, Winkler WG. 1987. Descriptive epidemiology from an epizootic of raccoon rabies in the middle Atlantic states, 1982-1983. *American Journal of Epidemiology* 126:429–37.

8. Krebs JW, Strine TW, Smith JS, Rupprecht CE, Childs JE. 1995. Rabies surveillance in the United States during 1994. *Journal of the American Veterinary Medical Association* 207:1562–75.

9. Heidt G, Ferguson D, Lammers J. 1982. A profile of reported skunk rabies in Arkansas: 1977–1979. *Journal of Wildlife Diseases* 18:269–77.